УДК 61:575 Вестник СПбГУ. Сер. 11. 2016. Вып. 1

В. М. Кочетков[1], *А. С. Глотов*[2], *Г. И. Образцова*[3]

АНАЛИЗ КОМБИНАЦИЙ ГЕНОТИПОВ УСЕЧЕННЫМ МЕТОДОМ ВАЛЬДА ДЛЯ ОПРЕДЕЛЕНИЯ РИСКА ЗАБОЛЕВАНИЯ

[1] Санкт-Петербургский государственный педиатрический медицинский университет, Российская Федерация, 194100, Санкт-Петербург, Литовская ул., 2

[2] НИИ акушерства, гинекологии и репродуктологии им. Д. О. Отта, Российская Федерация, 199034, Санкт-Петербург, Менделеевская линия, 3

[3] Федеральный центр сердца, крови и эндокринологии им. В. А. Алмазова, Российская Федерация, 197341, Санкт-Петербург, ул. Аккуратова, 2

Предложен метод оценки риска заболевания, основанный на выборе генотипов кандидатных генов, позволяющих выявить предрасположенность к заболеванию. Описана процедура принятия решения относительно принадлежности испытуемого пациента к группе риска. Приведен алгоритм оценки степени неоднородности таблиц с информацией о генотипах в группе риска и контрольной группе. Отмечены вычислительные ограничения метода и возможности его модификации для таблиц большого объема. Библиогр. 9 назв. Рис. 1.

Ключевые слова: кандидатные гены, анализ генотипов, последовательный анализ.

GENOTYPE COMBINATIONS ANALYSIS BY TRUNCATE WALD'S METHOD TO ESTIMATE THE DISEASE RISK

V. M. Kochetkov[1], *A. S. Glotov*[2], *G. I. Obraztsova*[3]

[1] St. Petersburg state Pediatric Medical University, 2, Litovskaya ul., St. Petersburg, 194100, Russian Federation

[2] The Research Institute of Obstetrics, Gynecology and Reproductology named after D. O. Ott, 3, Mendeleevskaya line, St. Petersburg, 199034, Russian Federation

[3] Federal Heart, Blood and endocrinology named after V. A. Almazov, 2, ul. Akkuratova, St. Petersburg, 197341, Russian Federation

On basis of candidate genes genotypes choice a method is proposed to reveal the susceptibility to disease. The procedure of decision on a patient belonging to risk group is described. An algorithm of heterogeneity degree evaluation for the tables containing information on genotypes of risk and control groups is considered. Some calculation restrictions of the method and possibilities of its modification for large tables are pointed to. Refs 9. Fig. 1.

Keywords: arterial hypertension, candidate genes, sequential analysis.

Введение

Поиску эффективных методов анализа генотипов на предмет выявления предрасположенности к различным заболеваниям в настоящее время посвящена обширная литература (см. [1–3] и представленную там библиографию). Основой этих методов является сравнение комбинаций генотипов кандидатных генов у представителей двух тестовых групп — группы риска и контрольной группы.

Один из традиционных подходов связан с использованием методов машинного обучения для поиска комбинаций тех генотипов, представленность которых различается в обеих указанных группах наиболее сильно, — это так называемый метод уменьшения мультифакториальной размерности [4].

В настоящей статье предлагается альтернативный подход, который может быть назван аналитическим, — он не использует алгоритмы машинного обучения, а базируется на принципе отношения правдоподобия и оценке вероятности безошибочного выбора гипотезы по методу А. Вальда [5]. Преимущества такого способа оценки предрасположенности к заболеванию, а также присущие ему ограничения будут видны из дальнейшего.

Подход, опирающийся на последовательный анализ Вальда, уже предлагался ранее [6] в связи с рассмотрением оценки риска детской гипертензии. В настоящей статье излагается его переработанная и существенно углубленная версия.

Суть решаемой задачи состоит в следующем. Пусть выявлены кандидатные гены, связанные с каким-либо заболеванием, и имеется набор данных, относящихся к пациентам группы риска и характеризующих значения генотипов кандидатных генов для каждого пациента этой группы. Набор этих данных назовем базовой таблицей T_1. Пусть аналогичная базовая таблица T_0 содержит такие же данные, относящиеся к пациентам контрольной группы.

Пусть число кандидатных генов, потенциально влияющих на риск заболевания, равно *g*. Генотипу каждого гена при этом может отвечать одно из трех состояний: гетерозигота и две различные гомозиготы. Для удобства обработки данных целесообразно кодировать гены и их генотипы цифрами или буквами алфавита, например способом, предложенным в [7].

Обе указанные таблицы имеют одинаковую структуру: число столбцов в них отвечает числу кандидатных генов *g*, а число строк — количеству обследованных пациентов в группе. В ячейках таблицы для каждого обследованного пациента проставляется код генотипа соответствующего гена.

После составления базовых таблиц целесообразно убедиться, что полный набор генотипов для пациента в одной из таблиц не повторен для какого-либо пациента в другой таблице. Если такой повтор имеет место, то соответствующие записи нужно удалить из обеих таблиц.

Предположим, что обследуется новый пациент, для которого получена информация относительно его генотипов, отвечающих кандидатным генам рассматриваемого заболевания. Соответствующий набор таких генотипов будем называть генным портретом пациента. Задача состоит в том, чтобы методами полигенного анализа определить степень риска соответствующего заболевания для пациента на основе данных, содержащихся в базовых таблицах.

В соответствии с тем, что в полигенном анализе учитывается одновременное действие нескольких генов, целесообразно рассматривать совокупно всевозможные комбинации генов и соответствующие им генотипы. Такие комбинации, содержащие коды генов и коды генотипов, будем называть цепочками.

Количество генов и их выбор в цепочке ничем не ограничивается, и цепочки считаются различными, если они отличаются хотя бы одним генотипом какого-то гена.

Цепочки генов можно строить на основе различных массивов данных:

1) исходя из имеющегося генного портрета пациента; такие цепочки будем далее называть *цепочками портрета*, их число равно $M_p = \sum_{i=1}^{g} C_g^i = 2^g - 1$,

2) на основе совокупности генотипов (возможно повторяющихся), содержащихся в строках базовых таблиц; такие цепочки назовем *реальными* — их количество $M_r = nM_p = n(2^g - 1)$, где n — объем таблицы,

3) исходя из абстрактного перебора всех возможных комбинаций кандидатных генов и их генотипов; такие цепочки будем далее называть *виртуальными*; их количество равно $M_v = \sum_{i=1}^{g} C_g^i 3^i = 4^g - 1$.

Нетрудно рассчитать, сколько раз конкретная цепочка встречается в одной из базовых таблиц. При известном числе обследованных пациентов в группе риска и контрольной группе это определит выборочную оценку вероятности, с которой данная цепочка включается в соответствующую в таблицу. Далее такую оценку вероятности будем для краткости называть долей.

Для рассмотрения статистики цепочек необходимо сформулировать некоторые исходные предположения. Они состоят в следующем.

1. Пусть число записей в базовой таблице равно n и для некоторого гена два из трех возможных генотипов проявлены в записях таблицы соответственно k_1 и k_2 раз (третий генотип проявлен, следовательно, $n–k_1–k_2$ раз). Предполагается, что величины k_1 и k_2 имеют полиномиальное распределение

$$P(k_1, k_2) = \frac{n!\, p_1^{k_1} p_2^{k_2} (1 - p_1 - p_2)^{n-k_1-k_2}}{k_1!\, k_2!\, (n - k_1 - k_2)!}, \qquad (1)$$

где p_1 и p_2 — вероятности реализации соответствующих генотипов ($p_1 + p_2 < 1$). При этих условиях число реализаций отдельного генотипа (без учета статистики прочих генотипов) имеет, как легко показать, биномиальное распределение.

2. Любая цепочка базовой таблицы, взятая как целое, имеет биномиальное распределение вероятности.

3. Цепочки в общем случае являются статистически зависимыми — вероятность реализации конкретной цепочки зависит от того, реализована ли для данного пациента другая цепочка.

В статистике для проблем, аналогичных задаче о принадлежности пациента к группе риска, обычно применяют термин дискриминация. В соответствии с этим на основе введенных терминов задачу определения риска заболевания можно определить так: по известному значению доли каждой цепочки в обеих исходных таблицах и по известному генному портрету пациента необходимо получить дискриминационное решение относительно принадлежности пациента к группе риска.

1. Алгоритм дискриминации

Перейдем к описанию алгоритма дискриминации риска заболевания. В настоящем разделе описываются лишь основы предлагаемого подхода к дискриминированию. Особые случаи, возникающие трудности и пути их преодоления рассматриваются в последующих разделах.

Для принятия или исключения гипотезы риска заболевания целесообразно использовать статистический критерий, обладающий наибольшей мощностью (о

мощности критерия см. [8]). Таким критерием, как известно, является отношение двух функций правдоподобия, каждая из которых отвечает одной из выбираемых гипотез.

В рассматриваемом случае функции правдоподобия могут строиться следующим образом. Выбирается одна или несколько цепочек, представленных в генном портрете пациента, и на основании базовых таблиц оценивается доля этих цепочек $L^{(1)}$ для группы риска и аналогичная величина $L^{(0)}$ для контрольной группы. Далее рассчитывается отношение правдоподобия $Z = L^{(1)} / L^{(0)}$, по величине которого принимается дискриминационное решение.

Удобной и экономной в отношении потребного объема информации разновидностью метода отношения правдоподобия является алгоритм последовательного анализа Вальда, когда пошагово учитываются новые данные, уточняющие отношение правдоподобия, и на каждом шаге оценивается возможность принятия одной из гипотез [5]. Именно этот подход и принят за основу последующих рассмотрений.

Дискриминация на основе метода Вальда может быть осуществлена следующим образом. По генному портрету пациента выбирается некоторая реальная цепочка (ей может, в частности, соответствовать отдельный ген), и для нее находятся доли $p_1^{(1)}$ и $p_1^{(0)}$, которыми она представлена в базовых таблицах.

Первый шаг процедуры Вальда состоит в расчете отношения $z_1 = p_1^{(1)} / p_1^{(0)}$. Далее берется другая цепочка портрета, для которой рассчитываются ее доли в базовых таблицах $p_{2|1}^{(1)}$ и $p_{2|1}^{(0)}$ при условии, что уже реализована первая цепочка. Рассчитывается новое отношение $z_2 = p_{2|1}^{(1)} / p_{2|1}^{(0)}$ и величина произведения $Z = z_1 z_2$. На следующем шаге выбирается еще одна цепочка портрета, для которой аналогично рассчитываются доли при условии, что реализованы первые две цепочки, и вычисляется соответствующее значение произведения $Z = z_1 z_2 z_3$, и т. д.

Пусть заданы допустимые вероятности ошибок α и β при выборе дискриминационного решения: α — вероятность ошибочно отвергнуть гипотезу риска и β — вероятность ее ошибочного принятия (так называемые ошибки первого и второго рода). Если на некотором шаге K величина

$$Z_K = \frac{p_1^{(1)} p_{2|1}^{(1)} p_{3|12}^{(1)} \cdots p_{K|1,2,\ldots K-1}^{(1)}}{p_1^{(0)} p_{2|1}^{(0)} p_{3|12}^{(0)} \cdots p_{K|1,2,\ldots K-1}^{(0)}} \qquad (2)$$

удовлетворяет перечисляемым ниже условиям, то процедуру следует закончить и получить искомое дискриминационное решение:

1. С достоверностью, определяемой значениями α и β, при $Z_K \le (1-\beta)/\alpha$ следует принять гипотезу риска. (3)

2. При $Z_K \le \beta/(1-\alpha)$ необходимо отвергнуть эту гипотезу, т. е. считать, что обследуемый пациент не относится к группе риска. (4)

В случае, когда ни одно из этих неравенств не выполнено, необходимо вычислить следующую величину Z_{K+1} и снова оценить возможность принятия решения, и т. д.

В отношении описанной процедуры необходимо отметить ее важную особенность. Если цепочки портрета на каждом шаге выбираются случайным образом, то реализуется классическая процедура последовательного анализа, при которой отсутствует априорная информация относительно возможного результата на последующем шаге. Однако в рассматриваемом случае доли всех цепочек портрета известны по базовым таблицам и это дает возможность выстроить последовательность цепочек таким образом, чтобы получить дискриминационное решение на возможно более раннем этапе, недостижимом для классической схемы.

Случайный отбор цепочек портрета следует считать нерациональным по следующей причине. Как показывает практика расчетов, при числе кандидатных генов порядка десяти и объемах выборки в базовых таблицах менее 500 после нескольких расчетных шагов информация в исходных таблицах может оказаться недостаточной, чтобы рассчитать требуемые условные доли, входящие в выражения для отношения правдоподобия (2). Выход состоит в том, чтобы по генному портрету пациента и записям в обеих базовых таблицах выбрать специальную цепочку, по которой дискриминационное решение сразу находится с определенным (и часто достаточным) уровнем достоверности уже после первого расчетного шага. Это оказывается важным, потому что при таком подходе исчезает необходимость расчета условных долей. Подобную специально отбираемую цепочку (алгоритм ее нахождения описывается ниже) назовем экстремальной цепочкой.

Важно отметить, что вероятность безошибочной дискриминации по экстремальной цепочке отвечает для рассматриваемого генного портрета (и имеющихся данных в базовых таблицах) предельно достижимому уровню достоверности, и этот уровень, в рамках рассматриваемого метода, не может быть улучшен учетом других цепочек. Таким образом, рассмотрение единственной цепочки оказывается достаточным для получения решения с точностью, обеспечиваемой исходными данными.

Описанную модификацию процедуры Вальда, когда решение принимается по одной цепочке, назовем усеченной (одношаговой) процедурой. Ниже дается краткое описание такой процедуры анализа.

2. Тестовая таблица. Одношаговая процедура анализа

Предположим, что базовая таблица, отвечающая контрольной группе, содержит n_0 записей, а базовая таблица, соответствующая группе риска, содержит n_1 записей. Пользуясь обеими таблицами, можно рассчитать количество представлений k_0 каждой реальной цепочки в таблице T_0, относящейся к контрольной группе, и аналогичное количество представлений k_1 для таблицы T_1 группы риска.

Пусть составлена более обширная таблица, называемая тестовой, которая включает *все возможные* виртуальные цепочки (их число равно $M_v = 4^g - 1$). Перенесем в эту таблицу рассчитанные числа представлений k_0 и k_1 цепочек, содержащихся в таблицах T_0 и T_1. Поскольку число реальных цепочек M_r меньше числа виртуальных цепочек M_v, то в тестовой таблице некоторые значения k_1 и k_0 окажутся нулевыми. Те цепочки, для которых нулевыми являются обе величины k_1 и k_0, из рассмотрения исключаются.

Будем предварительно считать, что рассматриваемые значения k_0 и k_1 положительны. Случаи нулевых значений одной из величин k_0 или k_1 рассматриваются далее в разделе 3.

Формирование тестовой таблицы завершим перестановкой ее записей в направлении убыли модуля логарифма величины $\lambda = k_1 n_0/(k_0 n_1)$, равной отношению долей цепочки в обеих базовых таблицах. Указанная величина $|\ln\lambda|$, как легко видеть, обладает следующими свойствами: 1) она монотонно растет при увеличении разницы долей цепочки в обеих базовых таблицах, 2) она не чувствительна к перестановке числителя и знаменателя в выражении для λ, в связи с чем не зависит от того, какая доля — k_0/n_0 или k_1/n_1 — оказывается превалирующей.

В отношении содержания тестовой таблицы следует подчеркнуть, что она включает в себя всю информацию обеих базовых таблиц (в форме чисел реализации реальных цепочек, т. е. значений k_0 и k_1) и, кроме того, для всех цепочек содержит величины $|\ln\lambda|$, играющие особую роль при осуществлении одношаговой процедуры дискриминации.

Реализация одношаговой процедуры предусматривает следующие действия.

Индицируется генный портрет пациента — выявляются генотипы g кандидатных генов.

Путем пошагового перемещения по тестовой таблице, находится цепочка с наибольшим значением $|\ln\lambda|$, представленная в генном портрете пациента. Такая цепочка и окажется экстремальной[1]. Если экстремальная цепочка не находится, то это означает, что данных в базовых таблицах недостаточно для принятия дискриминационного решения при выбранных значениях ошибок α и β.

В отношении оценки достоверности принимаемого дискриминационного решения можно опираться на соотношения (3) и (4), для которых величина Z_K соответствует одношаговому выбору. Однако можно получить и иную оценку с помощью анализа количества представлений экстремальной цепочки в базовых таблицах. Такая оценка может быть найдена на основании следующих соотношений.

Количество представлений m экстремальной цепочки в обеих базовых таблицах T_0 и T_1 следует биномиальному закону

$$P_0(m) = C_{n_0}^{m} p_0^{\,m} \left(1 - p_0\right)^{n_0 - m}, \quad P_1(m) = C_{n_1}^{m_1} p_1^{\,m} \left(1 - p_1\right)^{n_1 - m}, \tag{5}$$

где p_0 и p_1 — доли цепочки в базовых таблицах. Для $\lambda > 1$ из (5) можно найти целочисленные квантили m_0 и m_1, при которых величины $\sum_{m=m_0}^{n_0} P_0(m)$ и $\sum_{m=0}^{m_1} P_1(m)$ не превосходят задаваемых уровней ошибок α и β. Найденные квантили позволяют оценить достоверность принимаемого дискриминационного решения: при $m_0/n_0 < m_1/n_1$ вероятность безошибочной дискриминации будет не менее значения $1 - (\alpha + \beta)$.

В случае $\lambda < 1$ запись аналогичных формул представляется очевидной.

Следует отметить, что учет уровней ошибок по соотношениям (3–4) и (5) имеет различное содержание: в первом случае рассматривается последовательность

[1] Здесь излагается лишь схема метода, поэтому не учитывается вероятность множественных экстремальных цепочек, имеющих одну и ту же величину $|\ln\lambda|$ (и, возможно, разные знаки $\ln\lambda$).

различающихся цепочек, а во втором — единственная цепочка, представленная в обеих базовых таблицах.

В отношении описанного метода дискриминации не следует полагать, что имеется принципиальное различие между одношаговым и многошаговым подходами. Дело в том, что использование экстремальной цепочки является скрытым вариантом многошаговости, то есть фактически реализует обычную, неусеченную процедуру последовательного анализа на основе неявного поочередного учета генотипов для генов цепочки. Действительно, если оценивать доли цепочки, включающей K генов, т. е. рассчитывать вероятности $P(\Gamma_1\Gamma_2\ldots\Gamma_K)$ одновременной реализации K отдельных генотипов Γ_i в портрете пациента, то получим величину, в точности совпадающую с функцией правдоподобия в формуле (1). Этот факт является следствием последовательного выражения вероятности произведения зависимых событий через условную вероятность: $P(\Gamma_1\Gamma_2) = P(\Gamma_1)P(\Gamma_2|\Gamma_1)$.

Хотя по очередности шагов реализации описанная одношаговая процедура, строго говоря, не отвечает принципу последовательного анализа, но тем не менее отмеченная ее скрытая многошаговость позволяет считать эту процедуру вариантом усеченного метода Вальда.

Заметим, что после нахождения экстремальной цепочки учет состояний других, дополнительных генов не производится[2], поскольку, в соответствии со структурой тестовой таблицы, такому учету отвечало бы снижение величины $|\ln\lambda|$, т. е. уменьшение разницы в числе представлений цепочки для базовых таблиц T_0 и T_1. Таким образом, экстремальная цепочка реализует показатель дискриминации, оказывающийся наилучшим при имеющейся информации в базовых таблицах.

3. Оценка потенциальной доли цепочки при нулевом числе ее реализаций

Для обеих базовых таблиц высокая разница в числе представлений цепочки наблюдается тогда, когда экстремальная цепочка многократно представлена в одной из таблиц (T_0 или T_1), а в другой таблице ее доля равна нулю. Наличие нулевой выборочной доли не означает, что действительная вероятность представления p_0 для такой цепочки нулевая, — это значение может быть отличным от нуля, но при имеющемся объеме выборки реализовалось с вероятностью $(1 - p_0)^n$ число представлений, равное нулю.

Автор процедуры последовательного анализа [5] применительно к рассматриваемому случаю нулевой доли предложил закончить процедуру выбором решения в пользу той гипотезы, для которой функция правдоподобия не равна нулю. Однако в задаче полигенной дискриминации на основе одношаговой процедуры целесообразно найти для величины p_0 приближенную оценку в форме конечного значения.

Найти величину p_0 для индивидуальной цепочки не представляется возможным. Однако при определенных предположениях можно найти значение p_0, усредненное по совокупности имеющихся данных и относящееся, таким образом, ко всем цепочкам с нулевой реализацией для рассматриваемой базовой таблицы.

[2] Учет дополнительных генов может производиться лишь в целях контроля, см. раздел 5.

Пусть рассматривается одна из базовых таблиц T_0 или T_1 и пусть пространство элементарных событий составляют ее цепочки с известными весами $w_i = k_i/M_r = k_i/[n(2^g - 1)]$ и соответствующими долями, т. е. относительным количеством реализаций $p_i = k_i/n$. Весовые коэффициенты w_i очевидно удовлетворяют условию нормировки $\sum w_i = 1$.

Если для каких-то цепочек их доли совпадают, то эти цепочки следует объединить в конгломераты с соответствующим суммарным значением долей и весовых коэффициентов. В границах данного раздела подобный конгломерат для краткости будем по-прежнему именовать цепочкой.

Безусловная вероятность $P(p_i)$ того, что доля некоторой цепочки оценивается значением p_i, равна весовому коэффициенту w_i, а условная вероятность $P(0|p_i)$ того, что при известном значении p_i число реализаций цепочки будет нулевым, равна $1 - p_i$. Тогда по теореме Байеса находится значение «обратной» условной вероятности того, что потенциальная доля цепочки равна p_i при условии нулевой реализации:

$$P(p_i \mid 0) = \frac{P(0 \mid p_i)P(p_i)}{\sum\limits_{p_i} P(0 \mid p_i)P(p_i)}. \qquad (6)$$

В качестве искомой величины p_0 следует взять взвешенное с весами (6) значение доли p_i:

$$p_0 = \sum_{p_i} p_i P(p_i \mid 0) = \frac{\sum\limits_{p_i} p_i P(0 \mid p_i)P(p_i)}{\sum\limits_{p_i} P(0 \mid p_i)P(p_i)}. \qquad (7)$$

Суммирование в (6) и (7) производится по отдельным цепочкам, доли которых равны известным значениям p_i.

Как следует из описанной последовательности расчета, значения величины p_0 для обеих базовых таблиц оказываются, вообще говоря, различающимися. Эти значения определяют оценку параметра λ для экстремальных цепочек при нулевом числе реализаций таких цепочек в одной из базовых таблиц.

4. Критерий неоднородности базовых таблиц

Результат дискриминации не может иметь стопроцентную достоверность и достигается при известных значениях ошибок первого и второго рода. По этой причине целесообразно иметь инструмент, позволяющий до выполнения дискриминации оценивать степень статистического различия элементов базовых таблиц и, следовательно, — возможность использования таких таблиц для дискриминирования.

Сравнительная оценка таблиц производится следующим образом. Пусть раздельно по таблицам T_0 и T_1 для каждого кандидатного гена с номером i рассчитана величина $n_{ij}^{(t)}$ — количество реализаций генотипа с кодом j. Верхним индексом t, равным 0 или 1, отмечена используемая базовая таблица.

Построим двухстрочечную таблицу сопряженности признаков: в первой строке располагаются $3g$ элементов $n_{ij}^{(0)}$ $(1 \le i \le g,\ 1 \le j \le 3)$, относящиеся к таблице T_0, во

второй строке — аналогичные элементы для таблицы T_1. Для сумм табличных значений по строкам и столбцам примем соответственно обозначения $\overline{n}_{ij} = \sum_{t=0} n_{ij}^{(t)}$ и $\overline{n}^{(t)} = \sum_{i=1}^{g}\sum_{j=1}^{3} n_{ij}^{(t)}$.

Пусть по таблицам оценены вероятности реализации генотипов всех генов. Сохраняя принятую нотацию для гена, генотипа и таблицы, обозначим эти величины как $p_{ij}^{(t)}$. Если таблицы T_0 и T_1 статистически однородны, т. е. для них число реализаций генотипов всех генов следует полиномиальному распределению (1) с идентичными вероятностями реализации генотипов, то выполняются условия $p_{ij}^{(0)} = p_{ij}^{(1)}$. При этом величина

$$W = \left[\overline{n}^{(0)} + \overline{n}^{(1)}\right]\left\{\sum_{t=0}^{1}\sum_{i=1}^{g}\sum_{j=1}^{3}\left(\frac{\left[n_{ij}^{(t)}\right]^2}{\overline{n}_{ij}\cdot\overline{n}^{(t)}} - 1\right)\right\} \qquad (8)$$

при достаточно высоких $n_{0\,и}\, n_1$ имеет распределение хи-квадрат. Для оценки числа степеней свободы этого распределения следует учесть следующее:

1) полное число всех элементов таблицы равно $n_{tot} = 6g$;

2) параметрами, входящими в выражение для W и вычисляемыми по элементам таблицы, являются $\overline{n}_{i1}$ и $\overline{n}_{i2}$ $(1 \le i \le g)$. Величина $\overline{n}_{i3}$ не является отдельным вычисляемым параметром, поскольку для полиномиального распределения имеем $\overline{n}_{i3} = n_0 + n_1 - \overline{n}_{i1} - \overline{n}_{i2}$. Величина $\overline{n}^{(t)}$ также не является вычисляемым по таблице параметром, поскольку она выражается через объемы таблиц n_t $(t = 0,1)$, а именно: $\overline{n}^{(t)} = gn_t$. Общее число вычисляемых параметров равно, следовательно, $n_{calc} = 2g$;

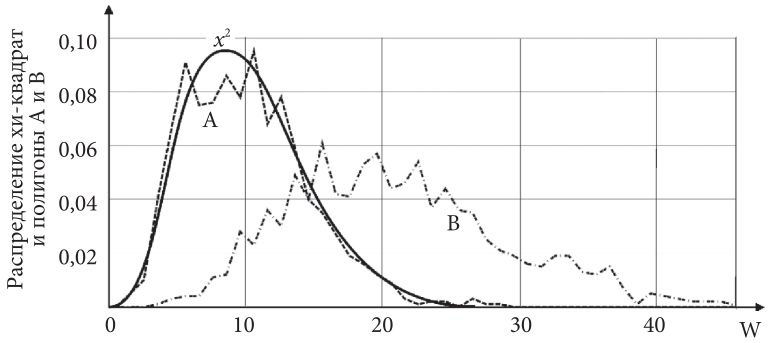
3) имеются условия, связывающие совокупности величин $n_{ij}^{(t)}$ с объемом таблиц T_0 и T_1: $\sum_{j=1}^{3} n_{ij}^{(t)} = n_t$, $(t = 0,1;\ i = 1,\ldots,g)$. Число таких условий равно $n_{cond} = 2g$.

Таким образом, число степеней свободы для распределения величины W выражается соотношением (см. [8])

$$df = n_{tot} - n_{calc} - n_{cond} = 2g.$$

Чем выше статистическое различие между элементами таблиц, выражающееся, в частности, в отличии вероятностей полиномиального распределения генотипов, тем большим становится математическое ожидание величины W по сравнению с его значением для распределения хи-квадрат. Это позволяет путем сравнения рассчитанного значения W и квантили распределения хи-квадрат делать общее заключение о пригодности базовых таблиц для выполнения дискриминирования.

С целью проверки описанного подхода были выполнены вычисления на основе построения 2000 модельных базовых таблиц объемом $n_0 = n_1 = 150$ (1000 таблиц T_0 и столько же таблиц T_1) при числе кандидатных генов $g = 5$.



Распределение хи-квадрат и полигоны распределения величины W при однородных (A) и неоднородных (B) базовых таблицах

Расчеты велись в двух режимах. В первом из них таблицы T_0 и T_1 были однородными, то есть для каждого гена вероятности реализаций генотипов p_1 и p_2 полиномиального распределения (1) выбирались одинаковыми, хотя они варьировались по случайному закону для разных генов с сохранением условия $p_1 + p_2 < 1$.

Для второго режима расчетов параметры полиномиального распределения p_1 и p_2 для таблиц T_0 и T_1 различались в пределах ± 20 %, что делало таблицы неоднородными.

На рисунке приведен вид распределения хи-квадрат с числом степеней свободы, равным $2g = 10$ (сплошная линия), и полигоны распределения величины (8), полученные для двух описанных режимов вычислений: линия A отвечает первому режиму вычислений (однородные таблицы), линия B — второму режиму (неоднородные таблицы). Из представленного рисунка видно, что для полигона A и распределения хи-квадрат основные параметры распределений — среднее и дисперсия — оказываются количественно близкими, что говорит о низком статистическом различии обоих распределений. Среднее же значение для полигона B примерно вдвое увеличено по сравнению со средним значением распределения хи-квадрат, имеющим величину $2g$.

Таким образом, описанный подход, основанный на сопоставлении рассчитанного значения величины W и квантили распределения хи-квадрат, позволяет количественно оценить пригодность базовых таблиц для выполнения дискриминирования.

Значение квантилей «распределения хи-квадрат» с достаточной точностью можно рассчитать по известным аппроксимациям [9], в частности по аппроксимации Корниша—Фишера.

Дополнительным средством проверки пригодности базовых таблиц для дискриминирования может служить анализ величин λ для тех цепочек тестовой таблицы, которые обладают наиболее высокими значениями $|\ln\lambda|$, — именно такие цепочки при дискриминации вероятнее всего окажутся экстремальными. Целесообразно убедиться, что для указанных цепочек удовлетворяются условия на достоверность принимаемого дискриминационного решения (см. соотношения (5) и описание использования квантилей биномиального распределения при оценке вероятности безошибочной дискриминации в разделе 3).

5. Вычислительные аспекты и улучшение метода при больших объемах выборки

Предложенный метод анализа генотипов, как и всякий вычислительный метод, использующий значительный объем исходной информации, имеет ограничения на допустимый объем входных данных. Следует отметить, что по времени вычислений наиболее затратной частью описанного метода является составление тестовой таблицы и ее заполнение на основе исходных базовых таблиц, относящихся соответственно к группе риска и контрольной группе.

Объем тестовой таблицы определяется полным количеством виртуальных цепочек, равным $4g - 1$, и при большом числе кандидатных генов g этот объем может оказаться значительным. Однако практика расчетов показывает, что при g порядка 10–11 проведение вычислений не встречает сложностей.

Тем не менее, при увеличении числа кандидатных генов могут возникать проблемы, однако и они частично могут решаться в направлении уменьшения числа анализируемых виртуальных цепочек на основе отбраковки тех цепочек, длина которых $0{,}8g$ и выше, если анализ показывает, что их частота в базовых таблицах невелика. Значимость длин цепочек для дискриминации на примере детской гипертензии отмечена в [6].

Следует учесть, что упомянутая наиболее длительная по времени процедура — создание тестовой таблицы — является разовым актом, после которого анализ любого количества генных портретов пациентов не требует сколько-нибудь значимых временных затрат.

Отметим, что предложенный метод дискриминации генотипов обеспечивает возможность оценки влияния отдельных кандидатных генов на величину риска заболевания. Для такой оценки по тестовой таблице анализируются цепочки с высокими значениями $|\ln\lambda|$ (например, не ниже 90% от максимальной величины), и для них определяется статистика включения всех кандидатных генов и их генотипов. В связи с тем, что значимость генов может иметь популяционные различия, последующая практика работы с генными портретами пациентов и выявление экстремальных цепочек могут обнаружить, что какими-то из кандидатынх генов для данной популяции можно пренебречь. Это позволит уменьшить в расчетах величину g и ускорить обработку данных.

Описанный метод дискриминации генотипов может быть значительно улучшен, если при числе кандидатных генов порядка 10 или менее обе базовые таблицы имеют большой объем — например, содержат 1000 и более записей. В этом случае целесообразно пересмотреть принципы построения статистики цепочек (см. раздел 1) и ориентироваться на более строгий подход. Этот подход содержит следующие уточнения.

Из-за статистической зависимости между цепочками не вполне верно считать, что в пределах какой-либо базовой таблицы любая цепочка имеет биномальное распределение, поскольку в разных строках таблицы подобная цепочка соседствует с различающимися наборами генотипов других цепочек.

Предположение о биномиальном распределении соответствует тому, что «фон влияния» на цепочку со стороны прочих цепочек усредняется. Однако такое усреднение можно не производить, и при больших объемах базовых таблиц возможен

более точный учет влияния на вероятность реализации цепочки со стороны генов, не включенных в цепочку.

Не входя в детали, укажем лишь направление описанного уточнения: реализация более точного подхода сопрягается с изучением корреляционных характеристик для комбинаций генотипов и последующим раздельным учетом идентичных генотипных комбинаций, если для них различаются их соседние корреляционно-значимые цепочки генотипов.

6. Выводы

1. Предложен аналитический метод анализа генотипов кандидатных генов на предмет оценки риска заболевания.

2. Метод может использоваться как независимо, так и параллельно с известными ранее способами анализа генотипов на основе принципов машинного обучения (MDR, RMDR, см. [4]). Одновременное использование аналитического метода и методов, ориентированных на машинное обучение, и последующее сравнение результатов, полученных обоими методами, могут повысить качество дискриминации.

3. По сравнению с методами MDR и RMDR предложенный аналитический метод обладает тем преимуществом, что позволяет количественно оценивать вероятность безошибочного дискриминационного решения.

4. Метод прошел тестирование как на модельных базовых таблицах, так и на реальных данных при числе кандидатных генов $g = 10$.

5. Предложенный метод позволяет выявлять в списке кандидатных генов те гены или комбинации генов, которые оказывают статистически наиболее значимое влияние на риск заболевания.

Литература

1. *Jiang Gui et al.* A robust multifactor dimensionality reduction method for detecting gene-gene interactions with application to the genetic analysis of bladder cancer susceptibility // Ann. Hum. Genet. 2011. Vol. 75(1). P. 20–28.

2. *Moore J. H, Williams S. M.* New strategies for identifying gene-gene interactions in hypertension // Ann. Med. 2002. Vol. 34. P. 88–95.

3. *Motsinger A. A., Ritchie M. D.* Multifactor dimensionality reduction: An analysis strategy for modelling and detecting gene — gene interactions in human genetics and pharmacogenomics studies // Human Genomics. 2006. Vol. 2. P. 318–328.

4. *Hahn L. W et al.* Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions // Bioinformatics. 2003. Vol. 19. P. 376–382.

5. *Вальд А.* Последовательный анализ. М.: ГИФМЛ, 1960. 328 с.

6. *Кочетков В. М.* и др. Дискриминация генотипов риска методом последовательного анализа Вальда для оценки предрасположенности детей к артериальной гипертензии // Вестн. С.-Петерб. ун-та. Сер. 11. 2015. Вып. 1. С. 35–45.

7. *Xiang W. et al.* BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies // Am. Journal of Human Genetics. Vol. 87. 2010. P. 325–340.

8. *Айвазян С. А., Мхитарян В. С.* Прикладная статистика и основы эконометрики**.** М.: Юнити, 1998. 1008 с.

9. *Goldstein R. B.* Chi-square quantiles, Algorithm 451 // Commun. Assoc. Comp. 1973. Vol. 16. P. 483–485.

References

1. Jiang Gui et al. A robust multifactor dimensionality reduction method for detecting gene-gene interactions with application to the genetic analysis of bladder cancer susceptibility. *Ann. Hum. Genet.*, 2011, vol. 75 (1), pp. 20–28.

2. Moore J. H, Williams S. M. New strategies for identifying gene-gene interactions in hypertension. *Ann. Med.*, 2002, vol. 34, pp. 88–95.

3. Motsinger A. A., Ritchie M. D. Multifactor dimensionality reduction: An analysis strategy for modelling and detecting gene — gene interactions in human genetics and pharmacogenomics studies. *Human Genomics*, 2006, vol. 2, pp. 318–328.

4. Hahn L. W et al. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics*, 2003, vol. 19, pp. 376–382.

5. Val'd A. *Posledovatel'nyj analiz* [*Sequential analysis*]. Moscow, GIFML Publ., 1960. 328 p. (In Russian)

6. Kochetkov V. M. i dr. Diskriminacija genotipov riska metodom posledovatel'nogo analiza Val'da dlja ocenki predraspolozhennosti detej k arterial'noj gipertenzii [перевод]. *Vestnik of Saint-Petersburg University. Ser. 11*, 2015, issue 1, pp. 35–45. (In Russian)

7. Xiang W. et al. BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am. Journal of Human Genetics*, vol. 87, 2010, pp. 325–340.

8. Ajvazjan S. A., Mhitarjan V. S. *Prikladnaja statistika i osnovy jekonometriki* [*Applied statistics and econometrics bases*]. Moscow, Unity, 1998. 1008 p. (In Russian)

9. Goldstein R. B. Chi-square quantiles, Algorithm 451. *Commun. Assoc. Comp.*, 1973, vol. 16, pp. 483–485.

Статья поступила в редакцию 28 января 2016 г.

Контактная информация:

Кочетков Валерий Михайлович — кандидат физико-математических наук, старший научный сотрудник; vmk@rzd-snw.ru
Образцова Галина Игоревна — доктор медицинских наук, доцент; galinaobraz@mail.ru
Глотов Андрей Сергеевич — кандидат биологических наук, старший научный сотрудник; anglotov@mail.ru

Kochetkov Valeriy M. — PhD, Senior Researcher; vmk@rzd-snw.ru
Obraztsova Galina I. — PhD, Associate Professor; galinaobraz@mail.ru
Glotov Andrey S. — PhD, Senior Researcher; anglotov@mail.ru